Sequence-based feature prediction and annotation of proteins

Juncker, Agnieszka; Jensen, Lars J.; Pierleoni, Andrea; Bernsel, Andreas; Tress, Michael L.; Bork, Peer; Von Heijne, Gunnar; Valencia, Alfonso; A Ouzounis, Christos; Casadio, Rita

Total number of authors:
11

Published in:
Genome Biology

Link to article, DOI:
10.1186/gb-2009-10-2-206

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Review

Sequence-based feature prediction and annotation of proteins

Agnieszka S Juncker*, Lars J Jensen[†], Andrea Pierleoni[‡], Andreas Bernsel[§], Michael L Tress[¶], Peer Bork[†], Gunnar von Heijne[§], Alfonso Valencia[¶], Christos A Ouzounis[¥], Rita Casadio[‡] and Søren Brunak*

Addresses: *Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark. [†]European Molecular Biology Laboratory, D-69117 Heidelberg, Germany. [‡]University of Bologna, Biocomputing Group, Via San Giacomo 9/2, 40126 Bologna, Italy. [§]Center for Biomembrane Research and Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden. [¶]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernández Almagro, 3, E-28029, Madrid, Spain. [¥]KCL Centre for Bioinformatics, School of Physical Sciences and Engineering, King's College London, London WC2R 2LS, UK.

Correspondence: Søren Brunak. Email: brunak@cbs.dtu.dk

Published: 2 February 2009

*Genome **Biology*** 2009, **10:**206 (doi:10.1186/gb-2009-10-2-206)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/2/206

Abstract

A recent trend in computational methods for annotation of protein function is that many prediction tools are combined in complex workflows and pipelines to facilitate the analysis of feature combinations, for example, the entire repertoire of kinase-binding motifs in the human proteome.

As more sequenced genomes become available, computational methods for predicting protein function from sequence data continue to be of high importance. In fact, such methods represent the only viable strategy for keeping up with the growth of genomic information. In the current era of pan- and metagenomics it is obvious that computational annotation is essential for turning sequence data into functional knowledge that can be used to understand biological mechanisms and their evolutionary trends.

From standalone function-prediction tools to workflows and pipelines

The computational annotation of structural and functional properties of proteins from their amino acid sequences is often possible, because similar functional or structural elements can be identified via similar sequence patterns. However, it is important to realize that there are two reasons for these similarities: some are due to homology (common ancestry), whereas others are due to convergent evolution (common selective pressure). This has consequences for the methods used to infer the annotations: while similarities due to common ancestry can often be identified by alignment techniques - either pairwise or profile-based - similarities produced by common selective pressures are often of a more subtle nature and are best identified using machine-learning techniques such as artificial neural networks, support vector machines (SVMs) or hidden Markov models adapted to the topology and sequential structure of the functional patterns in a given protein.

Functional patterns can be local, taking the shape of linear motifs or regions, or they can be reflected by more global features such as amino acid composition or pair frequencies, or by combinations of local and global features. Annotation based on homology has, in a broad sense, been used for as long as amino acid sequences have been compared. However, annotation of non-homologous patterns is also a very old discipline within bioinformatics. One of the very first published prediction methods in this context was a reduced-alphabet weight matrix calculating a score for signal peptide cleavage sites position by position [1].

No matter which type of functional feature a method attempts to identify, a crucial aspect of its usefulness is the predictive performance and, in particular, its ability to

generalize to novel, unannotated data [2]. The selection of dissimilar datasets for training, testing and validation is therefore critical to the practical usefulness of a given method. Overfitting to existing data has been and still is a common problem. When test and validation data are too similar to the training data, the predictive performance can be grossly overestimated or completely absent.

Interestingly, several of the breakthroughs in predicting functional features and structure have been linked to improvements in dataset preparation rather than to the invention of new algorithms as such [3-6]. Prediction of protein secondary structure represents one example [3,4], and of signal peptides another [6]. This also holds true for the new class of advanced workflow-oriented prediction schemes where hundreds of prediction tools are integrated [7]. The structuring of the experimental data and their conversion into datasets relevant for machine learning represents the most significant part of the inventive step, rather than the sophistication of the individual prediction tools [7].

In this review, we will provide an overview of how these different approaches can be used to annotate a number of functional features. We have chosen to focus on the structure-independent aspect of annotation - in other words, which features can be predicted without knowing or explicitly predicting the three-dimensional structure of the protein under consideration. Table 1 contains a list of web-sites with extensive references to such protein-annotation tools. We will begin by considering the identification of functionally important residues - that is, those involved in catalysis or binding. The prediction of post-translational modifications will be described - exemplified by phosphory-lation, glycosylation and lipid attachment. Then we will discuss how to predict which part of the cell a protein is destined for, on the basis of either the actual sorting signals or differences in global properties of proteins from different compartments. A related question is whether the protein is embedded in a membrane, and if so, which parts traverse the membrane and which parts are exposed to the two compart-ments separated by the membrane. Finally, we will discuss how these single-feature predictions can be integrated with each other and with overall homology-based detection schemes to assign a functional class to the entire protein.

An important current problem is to predict features that can be successfully used in comparative analysis of rather similar protein sequences, such as those derived from the same transcript by alternative splicing, from genome variation data (single-nucleotide polymorphisms, SNPs), variants arising by somatic mutation, or protein families from one or more species. Here the aim often is not to identify all func-tional features *per se*, but rather to single out differential functional features that may explain disease phenotypes or biochemical differences between organisms. The solution, as illustrated in Additional data file 1, is to structure and combine a large set of tools that can then be used to screen differential properties of datasets from large cohorts; this solution is now in development by the Epipe Consortium [8].

Table 1

Websites containing many references to popular protein annotation tools

http://www.bioinformatics.ca/links_directory
http://www.ncbi.nlm.nih.gov/Tools
http://www.ebi.ac.uk/Tools
http://www.expasy.org/tools
http://www.cbs.dtu.dk/services
http://hum-molgen.org/bioinformatics
http://sites.univ-provence.fr/~wabim/english/logligne.html
http://www.bioinformatics.fr/bioinformatics.php
http://www.brc.dcs.gla.ac.uk/~mallika/bioinformatics-tools.html

Some of these lists also contain references to data resources, but they all have special sections for prediction tools.

When many features are considered simultaneously, an effective way of structuring feature annotation is to develop an ontology of protein feature types. An ontology provides a structured and precisely defined common controlled vocabulary in a dynamic environment so that changes can occur as different uses are invented and new terms added. Recently, a new Protein Feature Ontology has been jointly developed by the BioSapiens, UniProt and Gene Ontology (GO) consortia [9], as an addition to the existing GO evidence ontology. This development is also very important for the future evolution of function-prediction tools.

Functional annotation of positional and non-positional features from sequence

While there often is a direct relationship between sequence similarity and conservation of protein structure, the same is not true for protein function: transfer of function based solely on the similarity between two sequences can be highly unreliable. Common evolutionary origin does not guarantee functional conservation of paralogs and the more distant the evolutionary relationship, the less reliable the transfer. Indeed, large-scale studies have shown that the transfer of functional annotation is only accurate for highly similar pairs of proteins [10,11]. However, even when two protein sequences do not appear to have overall sequence similarity, their alignment can contain short conserved sequence motifs, and these patterns of residues can be characteristic of a particular function. More powerful methods such as PSI-BLAST [12] or hidden Markov models can also be used to improve recognition performance. Methods such as ConFunc [13] and PFP [14] use clustering methods to refine and improve such homology-based predictions.

Domain databases such as Pfam [15], which recognizes the "accumulated sequence conservation of a long sequence segment" are also very useful tools for predicting function. Many Pfam functional domains and alignments are manually constructed by experts and are often among the best sources of functional information.

In many cases the most interesting functional information, such as catalytic and ligand-binding residues, is to be found at the residue level. One example of residue-level transfer can be found in the Catalytic Site Atlas [16]. Here catalytic residues extracted from the literature are supplemented by catalytic residues annotated from PSI-BLAST searches. One recent development has been Firestar [17], which is a server that integrates a database of experimentally validated functional residues with a sequence alignment analysis tool that evaluates the reliability of functional transfer. Firestar highlights potential functionally important residues such as ligand-binding residues and catalytic residues and allows users to assess whether the functionally important residues can be transferred.

Protein phosphorylation has a crucial role in almost all cellular signaling processes and is the most widespread post-translational modification in eukaryotes [18]. The first machine-learning-based method for prediction of phosphorylation sites, NetPhos, was published a decade ago; it uses ensembles of neural networks to distinguish between phosphorylated and non-phosphorylated residues [19].

However, mammals have more than 500 protein kinases with very different sequence specificities. Newer methods have thus instead focused on deriving separate sequence motifs for individual kinases or families of closely related kinases. The Scansite method relies on position-specific scoring matrices that are determined from data obtained in *in vitro* binding assays using degenerate peptide libraries [20]. Alternatively, machine-learning algorithms can be used to derive a sequence motif for each kinase (or kinase family) based on its known *in vivo* substrates. The first such method, NetPhosK, consisted of neural networks for only six kinase families [21], which later was extended to 17 families. Many other kinase-specifc methods have been developed using a variety of different machine-learning algorithms (see [22] and references therein for an overview).

As experimental phospho-proteomics approaches continue to produce vast numbers of phosphorylation sites, a key problem is to match these sites to the kinases that phosphorylate them. NetPhorest is a new atlas of consensus sequence motifs with a nonredundant collection of 125 sequence-based classifiers for linear motifs in phosphorylation-dependent signaling [23]. It covers more than 180 kinases and 100 phosphorylation-dependent binding domains (such as Src homology 2 (SH2), phosphotyrosine binding (PTB), BRCA1 C-terminal (BRCT), WW and 14-3-3). The resource is maintained by an automated pipeline, which uses phylogenetic trees to structure the available *in vivo* and *in vitro* data to derive probabilistic sequence models of linear motifs. This type of approach is therefore automatically maintained as new data become available and represents an entirely new angle on the sustainability of tools for protein function annotation.

The cellular substrate specificities of kinases are heavily influenced by contextual factors such as co-activators, protein scaffolds and expression [18]. The systems-biology-oriented method NetworKIN takes the context into account by augmenting the sequence motifs with a network context for the kinases and phosphoproteins [24]. The network is constructed on the basis of known and predicted functional associations from the STRING database, which integrates evidence from curated pathway databases, automatic literature mining, high-throughput experiments and genomic context [25]. For further details on prediction of biological networks see [26] and references therein.

Many proteins are glycoproteins and the most important types of glycosylations are N-linked, O-linked GalNAc (mucin-type), and O-β-linked GlcNAc (intracellular/nuclear) [21]. Glycosylation prediction is not a trivial task because of the lack of a clear consensus recognition sequence; however, it has been possible to develop useful models for prediction of O-GalNAc-glycosylation (NetOGlyc) using a neural network based approach that combines a range of features derived from sequence [27]. A recent advance in the glycosylation field has been the development of a new method - NetCGlyc - for predicting the unusual modification C-mannosylation [28].

Predicting subcellular localization

Automated sequence annotation of subcellular localization is a major step in protein functional annotation. This is particularly important in eukaryotic cells, which contain several subcellular compartments. Signal peptide prediction has a quite long history that will not be reviewed here. That area indeed represents one of the big successes in the entire field of predictive bioinformatics: algorithms are approaching a performance level comparable to the quality of the underlying experimental data, perhaps in some cases even better [6,29].

The SignalP scheme [30,31] was the first neural-network-based approach predicting both the presence of the secretory signal peptide and its cleavage site. It gave an order of magnitude improvement in performance. As mentioned above, this improvement was also based on new dataset preparation principles inspired by developments in protein structure prediction [4]. Other published machine-learning-based methods that perform well in this area include LOCTree [32], based on several binary SVMs, arranged in three different decision trees and specific for plants,

non-plants and prokaryotes; BaCelLo [29,33], which is based on a decision tree of binary SVMs, and is specific for animals, fungi and plants; TargetP [6], based on neural networks and specific for non-plants, plants and prokaryotes; WoLF PSORT [34], a classifier that computes a large number of sequence features and is specific for animals, fungi and plants. A general trend in the benchmarking of these algorithms is perhaps that the performance of multi-compartment predictors tends to be overestimated.

One subcellular location for which a wide range of sequence-based prediction methods has been developed is insertion into membranes. Structurally, integral membrane proteins come in two basic shapes, either tightly packed bundles of α-helices or β-barrels that often form permeable pores across the membrane. For various reasons, most computational work on membrane proteins has focused on the former. Generally speaking, topology predictors usually look for three important sequence characteristics of transmembrane alpha-helices: first, hydrophobic stretches of approximately 20 amino acids spanning the core of the lipid bilayer; second, a flanking 'aromatic belt' of tryptophan and tyrosine residues situated in the lipid-water interface; and third, an over-representation of the positively charged amino acids lysine and arginine in short cytoplasmic loops, known as the positive-inside rule [35].

Early attempts at predicting transmembrane topology from sequence were based on identifying peaks in hydrophobicity plots, using the positive-inside rule for uncertain cases and to predict the overall orientation of the protein [35]. More recent approaches use machine-learning algorithms to extract statistical sequence preferences from membrane proteins with known structures [36-40]. Including evolutionary information by basing the prediction on sequence profiles has been shown to increase performance levels by around 5-10% [37,39,41]. Current predictors attain around 80% accuracy on known membrane protein structures, although their performance might be overestimated when applied to whole-genome data [42].

In recent years, elucidation of the complexity of some membrane protein structures has led to the development of methods that predict not only transmembrane helices, but other structural features as well, such as re-entrant loops and interfacial helices [43,44]. Other methods, such as Phobius, combine the prediction of transmembrane helices with the simultaneous prediction of signal peptides, leading to improved performance levels for proteins that contain both [41].

A wide variety of proteins has been shown to contain covalently bound lipid groups [45]. Lipid anchor attachment is also a common way to link soluble proteins to membranes in eukaryotes. This modification directs the anchored protein to its very specific cellular location with an important impact on the final function. Predictors are presently available for modifications such as myristoylation, palmitoylation and prenylation [46,47]. The most common and best-studied lipid anchor modification is the glycosylphosphatidylinositol (GPI) linkage to the carboxy-terminal sequence portion that targets the protein toward the extracellular leaflet of the plasma membrane. In recent years, advances have also been made in predicting GPI-anchored proteins [48,49].

Global categories of biological function

Ultimately, the integration of various functional signals, ranging from key residues to signals for subcellular localization and post-translational modifications, can be extrapolated to global functional roles. These roles are typically expressed in general classification schemes, which aim at the complete description of known cellular functions of proteins [50]. Inspired by well-established catalogues, such as the Enzyme Committee (EC) nomenclature system for enzymes [51], these schemes comprise functional classes used in the characterization of genomes [52]. Similarly, generalized non-hierarchical structures, such as GO, express complex relationships between classes and subclasses [53]. One of the major challenges in function prediction is thus to capture the salient features of protein sequences and map those to existing functional classification schemes, often by combining information with other elements, for example subcellular localization or post-translational modifications.

Examples of this are represented by attempts to predict EC categories from sequence alone [54], the prediction of functional classes from keywords and other annotations [55], and finally the association of sequence with GO [56].

Non-homologous function prediction combining many features was first implemented in the ProtFun method for human proteins [57]. By design, the strength of the ProtFun method lies in classification of unannotated and orphan proteins. This strategy is based on the observation that proteins with the same function tend to exhibit similar feature patterns and functional similarity, which can be deduced from biochemical and biophysical properties such as average hydrophobicity, charge and amino acid composition as well as from local features such as glycosylation, phosphorylation and other post-translational modifications.

More recent methods have adopted a ProtFun-like approach in combination with homology or structural input and have reported improved performance, particularly in prediction of the GO categories [58,59]. One desirable element of function prediction is the association of annotation assignments to a score that reflects the quality of the assignment. The methods need to cluster the functional space into consistent clusters and subsequently provide probabilistic

estimates of assignment accuracy [60]; the recently developed method CORRIE can detect EC classes with high coverage [61]. Newer methods presumably benefit from the increasing quality and quantity of functional protein annotation. Furthermore, the combination of non-homologous prediction methods with homologous or structural methods is likely to overcome limitations inherent in each individual method.

A major challenge for the area of sequence-based protein function prediction is multi-functionality, where proteins have different roles in different compartments, tissues and organs. The low number of genes in the human genome has in itself increased the interest in experimental detection of this type of protein, and similarly, detection of alternative splicing by exon and tiling arrays also contributes large amounts of functional evidence of pleiotropy where a single gene influences multiple phenotypic traits. This situation calls for systems-biology-oriented approaches where data from protein interaction screens, gene expression data, and many other types of data are integrated. From a prediction perspective the entire area of multi-functional proteins is interesting as it also will call for new benchmarking principles for novel algorithms. Today most of the systems biology approaches still focus on proteins belonging to one single functional category. This problem indeed represents a major future challenge.

Additional Data Files

Additional data file 1 contains a workflow combining the prediction and annotation tools of the Epipe method and an example output.

References

1. von Heijne G: **Patterns of amino acids near signal-sequence cleavage sites.** *Eur J Biochem* 1983, **133:**17-21.
2. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview.** *Bioinformatics* 2000, **16:**412-424.
3. Hobohm U, Sander C: **Enlarged representative set of protein structures.** *Protein Sci* 1994, **3:**522-524.
4. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292:**195-202.
5. Nielsen H, Engelbrecht J, von Heijne G, Brunak S: **Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site.** *Proteins* 1996, **24:**165-177.
6. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools.** *Nat Protocols* 2007, **2:**953-971.
7. Miller ML, Jensen LJ, Diella F, Jørgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T, Olhovsky M, Pasculescu A, Alexander J, Knapp S, Blom N, Bork P, Li S, Cesareni G, Pawson T, Turk BE, Yaffe MB, Brunak S, Linding R: **Linear motif atlas for phosphorylation-dependent signaling.** *Sci Signal* 2008, **1:**ra2.
8. **EPipe 1.0** [http://www.cbs.dtu.dk/services/EPipe]
9. Reeves GA, Eilbeck K, Magrane M, O'Donovan C, Montecchi-Palazzi L, Harris MA, Orchard S, Jimenez RC, Prlic A, Hubbard TJ, Hermjakob H, Thornton JM. **The Protein Feature Ontology: a tool for the unification of protein feature annotations.** *Bioinformatics* 2008, **24:**2767-2772.
10. Devos D, Valencia A: **Practical limits of function prediction.** Proteins 2000, **41:**98-107.
11. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307:**1113-1143.
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
13. Wass MN, Sternberg MJ: **ConFunc - functional annotation in the twilight zone.** *Bioinformatics* 2008, **24:**798-806.
14. Hawkins T, Luban S, Kihara D: **Enhanced automated function prediction using distantly related sequences and contextual association by PFP.** *Protein Sci* 2006, **15:**1550-1556.
15. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36(Database issue):**D281-D288.
16. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32(Database issue):** D129-D133.
17. Lopez G, Valencia A, Tress ML. **Firestar - prediction of functionally important residues using structural templates and alignment reliability.** *Nucleic Acids Res* 2007, **35(Web Server issue):**W573-W577.
18. Ubersax JA, Ferrell JE Jr: **Mechanisms of specificity in protein phosphorylation.** *Nat Rev Mol Cell Biol* 2007, **8:**530-541.
19. Blom N, Gammeltoft S, Brunak S: **Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites.** *J Mol Biol* 1999, **294:**1351-1362.
20. Obenauer JC, Cantley LC, Yaffe MB: **Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.** *Nucleic Acids Res* 2003, **31:**3635-3641.
21. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S: **Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence.** *Proteomics* 2004, **4:**1633-1649.
22. Wan J, Kang S, Tang C, Yan J, Ren Y, Liu J, Gao X, Banerjee A, Ellis LB, Li T: **Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection.** *Nucleic Acids Res* 2008, **36:**e22.
23. Miller ML, Jensen LJ, Diella F, Jørgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T, Olhovsky M, Pasculescu A, Alexander J, Knapp S, Blom N, Bork P, Li S, Cesareni G, Pawson T, Turk BE, Yaffe MB, Brunak S, Linding R: **Linear Motif Atlas for phosphorylation-dependent signaling.** *Sci Signal* 2008, **1:**ra2.
24. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jørgensen C, Miron IM, Diella F, Colwill K, Taylor L, Elder K, Metalnikov P, Nguyen V, Pasculescu A, Jin J, Park JG, Samson LD, Woodgett JR, Russell RB, Bork P, Yaffe MB, Pawson T: **Systematic discovery of *in vivo* phosphorylation networks.** *Cell* 2007, **129:**1415-1426.
25. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P: **STRING 7 recent developments in the integration and prediction of protein interactions.** *Nucleic Acids Res* 2007, **35(Database issue):**D358-D362.
26. Harrington ED, Jensen LJ, Bork P: **Predicting biological networks from genomic data.** *FEBS Lett* 2008, **582:**1251-1258.
27. Julenius K, Mølgaard A, Gupta R, Brunak S: **Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites.** *Glycobiology* 2005, **15:**153-164.
28. Julenius K: **NetCGlyc 1.0: prediction of mammalian C-mannosylation sites.** *Glycobiology* 2007, **17:**868-876.
29. Pierleoni A, Martelli PL, Fariselli P, Casadio R: **BaCelLo: a balanced subcellular localization predictor.** *Nat Protocols Network* (DOI:10.1038/nprot.2007.165).
30. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340:**783-795.
31. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10:**1-6.
32. Nair R, Rost B: **Sequence conserved for subcellular localization.** *Protein Sci* 2002, **11:**2836-2847.
33. Pierleoni A, Martelli PL, Fariselli P, Casadio R: **BaCelLo: a balanced subcellular localization predictor.** *Bioinformatics* 2006, **22:**e408-e416.
34. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic Acids Res* 2007, **35(Web server issue):**W585-W587.

35. von Heijne G: **Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule.** *J Mol Biol* 1992, **225:**487-494.
36. Krogh A. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.
37. Jones DT: **Improving the accuracy of transmembrane protein topology prediction using evolutionary information.** *Bioinformatics* 2007, **23:**538-544.
38. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server.** *Bioinformatics* 2001, **17:**849-850.
39. Viklund H, Elofsson A: **Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information.** *Protein Sci* 2004, **13:**1908-1917.
40. Amico M, Finelli M, Rossi I, Zauli A, Elofsson A, Viklund H, von Heijne G, Jones D, Krogh A, Fariselli P, Martelli PL, Casadio R: **PONGO: a web server for multiple predictions of all-alpha transmembrane proteins**. *Nucleic Acids Res* 2006, **34(Web server issue):**169-172.
41. Käll L, Krogh A, Sonnhammer EL: **An HMM posterior decoder for sequence feature prediction that includes homology information.** *Bioinformatics* 2005, **21(Suppl 1):**i251-i257.
42. Melen K, Krogh A, von Heijne G: **Reliability measures for membrane protein topology prediction algorithms.** *J Mol Biol* 2003, **327:**735-744.
43. Viklund H, Granseth E, Elofsson A: **Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes.** *J Mol Biol* 2006, **361:**591-603.
44. Lasso G, Antoniw JF, Mullins JG: **A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops.** *Bioinformatics* 2006, **22:**e290-e297.
45. Resh MD: **Trafficking and signalling by fatty-acylated and prenylated proteins.** *Nat Chem Biol* 2006, **2:**584-590.
46. Zhou F, Xue Y, Yao X, Xu Y: **CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS).** *Bioinformatics* 2007, **22:**894-896.
47. Eisenhaber B, Eisenhaber F: **Post-translational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure?** *Curr Protein Pept Sci* 2007, **8:**197-203.
48. Poisson G, Chauve C, Chen X, Bergeron A: **FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring.** *Genomics Proteomics Bioinformatics* 2007, **5:**121-130.
49. Pierleoni A, Martelli PL, Casadio R. Pierleoni A, Martelli PL, Casadio R: **PredGPI: a GPI-anchor predictor.** *BMC Bioinformatics* 2008, **9:** 392.
50. Ouzounis CA, Coulson RM, Enright AJ, Kunin V, Pereira-Leal JB: **Classification schemes for protein structure and function.** *Nat Rev Genet* 2003, **4:**508-519.
51. Tipton K, Boyce S: **History of the enzyme nomenclature system.** *Bioinformatics* 2000, **16:**34-40.
52. Riley M: **Systems for categorizing functions of gene products.** *Curr Opin Struct Biol* 1998, **8:**388-392.
53. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
54. des Jardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA: **Prediction of enzyme classification from protein sequence without the use of sequence similarity.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**92-99.
55. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A: **EUCLID: automatic classification of proteins in functional classes by their database annotations.** *Bioinformatics* 1998, **14:**542-543.
56. Jensen LJ, Gupta R, Staerfeldt HH, Brunak S: **Prediction of human protein function according to Gene Ontology categories.** *Bioinformatics* 2003, **19:**635-642.
57. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Stærfeldt H, Rapacki K, Workman C, Andersen CAF, Knudsen S, Krogh A, Valencia A, Brunak S: **Prediction of human protein function from post-translational modifications and localization features.** *J Mol Biol* 2002, **319:**1257-1260.
58. Pal D, Eisenberg D: **Inference of protein function from protein structure.** *Structure* 2005, **13:**121-130.
59. Lobley AE, Nugent T, Orengo CA, Jones DT: **FFPred: an integrated feature-based function prediction server for vertebrate proteomes.** *Nucleic Acids Res* 2008, **36(Web Server issue):**W297-W302.
60. Levy ED, Ouzounis CA, Gilks WR, Audit B: **Probabilistic annotation of protein sequences based on functional classifications.** *BMC Bioinformatics* 2005, **6:**302.
61. Audit B, Levy ED, Gilks WR, Goldovsky L, Ouzounis CA: **CORRIE: enzyme sequence annotation with confidence estimates.** *BMC Bioinformatics* 2007, **8(Suppl 4):**S3.